PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-167581
(43)Date of publication of application : 22.06.1999

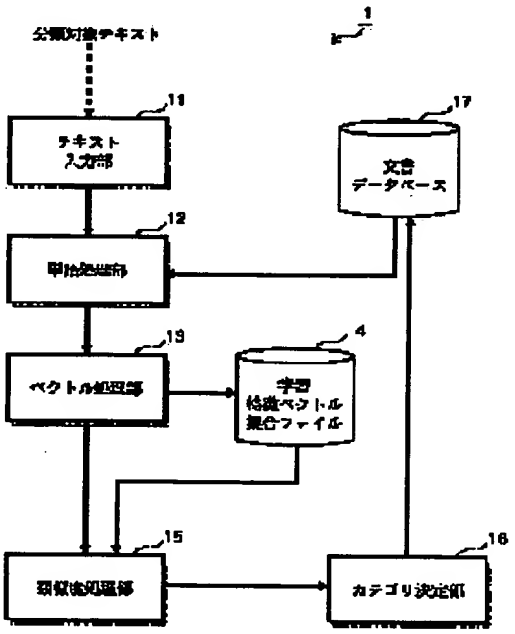
(51)Int.Cl. G06F 17/30
G06F 7/24

(21)Application number : 09-334309
(22)Date of filing : 04.12.1997

(71)Applicant : NTT DATA CORP
(72)Inventor : HARA MASAMI
KITANI TSUYOSHI

(54) INFORMATION SORTING METHOD, DEVICE AND SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide an information sorting device which can sort the texts with high accuracy.
SOLUTION: An information sorting device 1 includes a text input part 11, a word processing part 12, a vector processing part 13, a learning feature vector set file 14, a similarity processing part 15, a category decision part 16 and an external or internal document data base 17. The part 12 calculates the importance of category of every word that is extracted from a learning text based on both number of appearance and category frequencies of the word. The part 15 calculates the similarity of words based on the learning feature vector, the learning feature vector set and the sorting object text feature vector which are calculated based on the importance of words calculated at the part 12. The part 16 decides a prescribed number of corresponding categories as the categories of the sorting object texts based on the similarity having the largest calculation value. Then the sorting object texts sorted in each category are stored in the" data base 17.



LEGAL STATUS

[Date of request for examination] 06.06.2000

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平11－167581

(43)公開日　平成11年（1999）6月22日

(51)Int.Cl.[6]	識別記号	FI		
G06F 17/30		G06F 15/401	310	D
7/24		7/24		A
		15/40	370	A
		15/403	350	C

審査請求　未請求　請求項の数11　ＯＬ（全 8 頁）

(21)出願番号　特願平9－334309

(22)出願日　平成9年（1997）12月4日

(71)出願人　000102728
株式会社エヌ・ティ・ティ・データ
東京都江東区豊洲三丁目3番3号

(72)発明者　原　正巳
東京都江東区豊洲三丁目3番3号　エヌ・ティ・ティ・データ通信株式会社内

(72)発明者　木谷　強
東京都江東区豊洲三丁目3番3号　エヌ・ティ・ティ・データ通信株式会社内

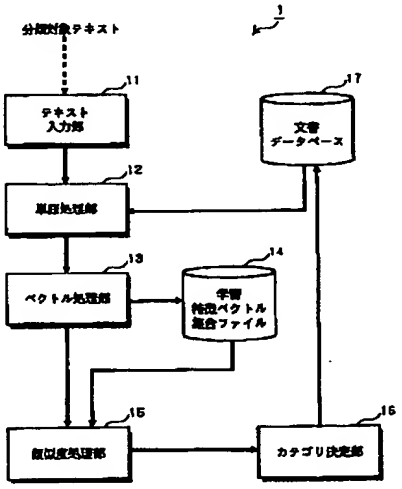
(74)代理人　弁理士　鈴木　正剛

(54)【発明の名称】情報分類方法、装置及びシステム

(57)【要約】

【課題】　高精度なテキスト分類が可能な情報分類装置を提供する。

【解決手段】　テキスト入力部１１、単語処理部１２、ベクトル処理部１３、学習特徴ベクトル集合ファイル１４、類似度処理部１５、カテゴリ決定部１６を備え、外部または内部に文書データベース１７を具備して情報分類装置１を構成する。単語処理部１２では、学習テキストから抽出された各単語のカテゴリに対する重要度を単語の出現件数及びカテゴリ頻度に基づいて算出する。類似度処理部１５では、当該重要度に基づいて算出された学習特徴ベクトル及び学習特徴ベクトル集合と分類対象テキストの特徴ベクトルとに基づいて類似度を算出する。カテゴリ決定部１５では、算出値が最大となる類似度から所定数の対応カテゴリを分類対象テキストのカテゴリとして決定し、このカテゴリによって分類された分類対象テキストが文書データベース１７に蓄積されるようにする。



1

【特許請求の範囲】

【請求項１】　属すべきカテゴリが既知の学習用テキストから単語を抽出し、抽出した単語毎に、その出現件数及び出現するカテゴリ数に基づく重要度を算出するとともに、算出された重要度を要素としてカテゴリ毎の特徴を表す学習特徴ベクトルを生成する過程と、

カテゴリが不明な分類対象テキストに対して当該分類対象テキスト中の単語毎の出現頻度に基づく重要度を算出し、算出された重要度を要素としてテキスト毎の特徴を表す分類対象特徴ベクトルを生成する過程と、

分類対象特徴ベクトルと前記カテゴリ毎の学習特徴ベクトルとの類似度を判定する過程とを含み、

前記分類対象テキストとの類似度が所定範囲内の学習特徴ベクトルに対応するカテゴリを当該分類対象テキストに付与すべきカテゴリ候補とすることを特徴とする情報分類方法。

【請求項２】　前記分類対象テキストとの類似度の高い順に並べたときに上位から予め定めた件数以上の学習特徴ベクトルに対応するカテゴリを当該分類対象テキストに付与すべきカテゴリ候補とすることを特徴とする請求項１記載の情報分類方法。

【請求項３】　前記学習特徴ベクトルを生成する過程は、

前記学習用テキスト中の単語の出現傾向に着目してカテゴリの特徴を表す指標となる特徴語及びカテゴリに依存しない一般語を判別し、前記単語の出現するカテゴリ数に基づいて前記一般語の重要度を低減させることで前記特徴語の重要度が相対的に高く反映された学習特徴ベクトルを生成することを特徴とする請求項１記載の情報分類方法。

【請求項４】　１または複数のカテゴリが付与された学習用テキストの分類体系に即してカテゴリが不明な分類対象テキストに付与すべきカテゴリを決定して分類処理を行う装置であって、

前記学習用テキスト及び分類対象テキストの各々から単語を抽出するとともに抽出した単語毎の重要度を算出する単語処理手段と、

前記単語毎の重要度を要素として、前記学習用テキストの特徴をカテゴリ毎に表現した学習特徴ベクトル、及び分類対象テキストの特徴をテキスト毎に表現した分類対象特徴ベクトルを生成するベクトル処理手段と、

個々の分類対象特徴ベクトルと前記学習特徴ベクトルとの特徴差に基づいてカテゴリ毎の学習特徴ベクトルに対する前記分類対象特徴ベクトルの類似度を判定する類似度処理手段と、

前記類似度処理手段による判定結果に基づいて、前記分類対象テキストに付与すべきカテゴリを決定するカテゴリ決定手段と、

を備えることを特徴とする情報分類装置。

【請求項５】　前記単語処理手段は、前記学習用テキスト中の総カテゴリ数を特定の単語が出現するカテゴリ数による除算に基づくカテゴリ頻度係数を算出する手段を有し、

特定のカテゴリに出現する単語の出現件数と前記カテゴリ頻度係数との乗算により前記学習用テキスト中の単語毎の重要度を算出するとともに、出現件数が相対的に多く且つカテゴリへの依存が相対的に少ない単語の重要度を低減させるように構成されていることを特徴とする請求項４記載の情報分類装置。

【請求項６】　前記単語処理手段は、特定のカテゴリに出現する単語の出現件数と前記カテゴリ頻度係数との乗算による算出値に、さらに当該単語の出現頻度を乗算することにより前記学習用テキスト中の単語毎の重要度を算出するように構成されていることを特徴とする請求項４記載の情報分類装置。

【請求項７】　前記単語処理手段は、前記分類対象テキスト中の単語の出現頻度を計測する手段を有し、出現頻度が低い単語ほど当該分類対象テキスト中の重要度が高くするように構成されていることを特徴とする請求項４記載の情報分類装置。

【請求項８】　前記類似度処理手段は、個々の学習特徴ベクトル及び分類対象特徴ベクトル間の内積に基づいて両ベクトルの余弦を算出するとともに、この余弦の算出値を所定順に整列して両ベクトルの特徴差を定量化するように構成されていることを特徴とする請求項４記載の情報分類装置。

【請求項９】　前記分類対象テキストに対する類似度が所定範囲内となる１または複数の学習特徴ベクトルに対応するカテゴリを視認可能にして提示する提示手段をさらに備え、

前記カテゴリ決定手段は、前記提示手段による提示に対応して特定されたカテゴリを当該分類対象テキストに付与すべきカテゴリとして決定するように構成されていることを特徴とする請求項４記載の情報分類装置。

【請求項１０】　請求項４ないし９のいずれかの項に記載された情報分類装置と、通信回線を介して流通する前記分類対象テキストを前記情報分類装置に取り込むテキスト入力手段とを備えたことを特徴とする情報分類システム。

【請求項１１】　前記テキスト入力手段は、前記分類対象テキストをエージェント機能を通じて前記情報分類装置に入力するように構成されていることを特徴とする情報分類システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、自然言語処理や情報検索技術分野において、電子化されたテキスト群を効率的に分類する情報分類手法に関する。

【０００２】

【従来の技術】電子化情報の分類手法として、ベクトル

表現したカテゴリの特徴と未分類の電子化情報とを比較することによりカテゴリを決定する手法が知られている。以下、この手法の概要を説明する。前提条件として、カテゴリの特徴を抽出するために利用する学習用電子化情報（以下、学習テキスト）群には、予めN種類のカテゴリＣ１，Ｃ２，…、ＣＮが各々付与されているものとする。

【０００３】まず、カテゴリＣｉ（１≦ｉ≦Ｎ）の特徴を表す特徴ベクトルｐｉを作成するために、カテゴリＣｉが付与されている学習テキスト群から単語を抽出する。そして、学習テキストにおける各単語の重要度をカテゴリ毎に決定する。重要度の決定方法としては、情報検索の分野で提案されたＴＦ・ＩＤＦ法が広く知られている（「Introduction to Modern Information Retrieval：G.Salton著、McGraw-Hill」参照）。このＴＦ・ＩＤＦ法における単語の重要度は、出現頻度ｔｆと、出現件数ｄｆの逆数ｉｄｆとを用いて定義される。具体的には、カテゴリＣｉにおける単語ｔｋの重要度Ｗ（ｔｋ，Ｃｉ）は、以下に示す式（１）で算出される。

$$W(tk,Ci)=tf(tk,Ci)\log(Li/df(tk,Ci)+1) \quad \cdots(1)$$

ここでｔｆ（ｔｋ，Ｃｉ）は、カテゴリＣｉにおける単語ｔｋの出現頻度を表し、またｄｆ（ｔｋ，Ｃｉ）は、カテゴリＣｉにおける単語ｔｋの出現件数を表している。一方、Ｌｉは、カテゴリＣｉにおける総テキスト件数を表している。

【０００４】次に、学習テキスト集合に出現するすべての単語ｔ１，ｔ２，～，ｔＭについて上記式（１）によりカテゴリＣｉにおける重要度を各々算出し、算出された各重要度を要素としたベクトルをカテゴリＣｉの特徴ベクトルｐｉとする。未分類テキストＴについても同様に、特徴ベクトルｑを算出する。この場合の特徴ベクトルの要素となる単語の重要度には、主に出現頻度ｔｆが用いられる。未分類テキストＴにおけるカテゴリの決定には、各カテゴリの特徴ベクトルｐｉ（１≦ｉ≦Ｎ）と未分類テキストＴの特徴ベクトルｑとの類似度ｄ（ｐｉ，ｑ）が用いられる。この類似度計算の代表的な例には、両ベクトルの内積を算出する方法や集合論的測度を利用する方法等が知られており、「情報検索：伊藤哲朗著、昭晃堂」に詳しく記述されている。

【０００５】このように、カテゴリ毎に上述の類似度ｄ（ｐｉ，ｑ）を算出して利用することにより、未分類テキストＴと類似の度合いが近いカテゴリを複数選択して分類先となるカテゴリを決定する。

【０００６】

【発明が解決しようとする課題】上述のように、ＴＦ・ＩＤＦ法は、例えば検索語と検索データベース内のテキストとを比較するためのベクトル作成に利用される手法であり、出現頻度ｔｆが大きいほど出現件数の逆数ｉｄｆが大きい、即ち出現件数ｄｆが小さいほど重要度が高くなるものである。

【０００７】しかし、テキストの分類では、ベクトル作成の対象となるテキスト群は、通常、同一カテゴリに属しており、カテゴリを考慮しない情報検索とはテキストの特徴が異なったものとなる。そのため、カテゴリの特徴となるような重要な単語（以下、特徴語）は、同一カテゴリに属するテキストに着目した場合には、多くのテキストに出現する、即ち出現件数ｄｆが大きいことが考えられる。このことは、出現件数ｄｆの逆数を用いたｉｄｆを利用するＴＦ・ＩＤＦ法では、特徴語に低い重要度を付与してしまう可能性があることを意味する。この結果、ＴＦ・ＩＤＦ法を利用して単語の重要度を決定すると、カテゴリの特徴を明確に表現した特徴ベクトルの作成が困難となり、また、分類精度も低下してしまうという問題があった。

【０００８】一方、出現件数ｄｆを利用する場合でも、出現件数ｄｆの多い単語群にはカテゴリに依存することなく出現する一般的な語（以下、一般語）も含まれており、出現件数ｄｆが多い語が必ずしも特徴語であるとはいえない。このため、特徴語の重要度に出現件数ｄｆ自体が利用されることは殆どなかった。

【０００９】そこで、本発明の課題は、学習テキストにおけるカテゴリの特徴語となる単語の重要度を考慮することにより、高精度の分類を可能にする新規な情報分類方法を提供することにある。また、本発明の他の課題は、上記情報分類方法の実施に適した情報分類装置、及び情報分類システムを提供することにある。

【００１０】

【課題を解決するための手段】上記課題を解決するため、本発明は、属すべきカテゴリが既知の学習用テキストから単語を抽出し、抽出した単語毎に、その出現件数及び出現するカテゴリ数に基づく重要度を算出するとともに、算出された重要度を要素としてカテゴリ毎の特徴を表す学習特徴ベクトルを生成する過程と、カテゴリが不明な分類対象テキストに対して当該分類対象テキスト中の単語毎の出現頻度に基づく重要度を算出し、算出された重要度を要素としてテキスト毎の特徴を表す分類対象特徴ベクトルを生成する過程と、分類対象特徴ベクトルと前記カテゴリ毎の学習特徴ベクトルとの類似度を判定する過程とを含み、前記分類対象テキストとの類似度が所定範囲内の学習特徴ベクトル、または類似度の高い順に並べたときに上位から予め定めた件数以上の学習特徴ベクトルに対応するカテゴリを当該分類対象テキストに付与すべきカテゴリ候補とする、情報分類方法を提供する。

【００１１】この情報分類方法において、前記学習特徴ベクトルを生成する過程は、例えば、前記学習用テキスト中の単語の出現傾向に着目してカテゴリの特徴を表す指標となる特徴語及びカテゴリに依存しない一般語を判別し、前記単語の出現するカテゴリ数に基づいて前記一般語の重要度を低減させることで前記特徴語の重要度が

相対的に高く反映された学習特徴ベクトルを生成することを特徴とする。

【００１２】上記他の課題を解決する本発明の情報分類装置は、１または複数のカテゴリが付与された学習用テキストの分類体系に即してカテゴリが不明な分類対象テキストに付与すべきカテゴリを決定して分類処理を行う装置であって、以下の要素を備えて構成される。

（１）前記学習用テキスト及び分類対象テキストの各々から単語を抽出するとともに抽出した単語毎の重要度を算出する単語処理手段。この単語処理手段は、例えば、前記学習用テキスト中の総カテゴリ数を特定の単語が出現するカテゴリ数による除算に基づくカテゴリ頻度係数を算出する手段を有し、特定のカテゴリに出現する単語の出現件数と前記カテゴリ頻度係数との乗算により前記学習用テキスト中の単語毎の重要度を算出するとともに、出現件数が相対的に多く且つカテゴリへの依存が相対的に少ない単語の重要度を低減させるように構成される。また、特定のカテゴリに出現する単語の出現件数と前記カテゴリ頻度係数との乗算による算出値に、さらに当該単語の出現頻度を乗算することにより前記学習用テキスト中の単語毎の重要度を算出するように構成される。あるいは、前記分類対象テキスト中の単語の出現頻度を計測する手段を有し、出現頻度が低い単語ほど当該分類対象テキスト中の重要度が高くするように構成される。

（２）前記単語毎の重要度を要素として、前記学習用テキストの特徴をカテゴリ毎に表現した学習特徴ベクトル、及び分類対象テキストの特徴をテキスト毎に表現した分類対象特徴ベクトルを生成するベクトル処理手段。

（３）個々の分類対象特徴ベクトルと前記学習特徴ベクトルとの特徴差に基づいてカテゴリ毎の学習特徴ベクトルに対する前記分類対象特徴ベクトルの類似度を判定する類似度処理手段。この類似度処理手段は、例えば、個々の学習特徴ベクトル及び分類対象特徴ベクトル間の内積に基づいて両ベクトルの余弦を算出するとともに、この余弦の算出値を所定順に整列して両ベクトルの特徴差を定量化するように構成される。

（４）前記類似度処理手段による判定結果に基づいて、前記分類対象テキストに付与すべきカテゴリを決定するカテゴリ決定手段。

【００１３】好ましくは、前記分類対象テキストに対する類似度が所定範囲内となる１または複数の学習特徴ベクトルに対応するカテゴリを視認可能にして提示する提示手段をさらに備える。この場合、前記カテゴリ決定手段は、前記提示手段による提示に対応して特定されたカテゴリを当該分類対象テキストに付与すべきカテゴリとして決定するように構成する。

【００１４】上記他の課題を解決する本発明の情報分類システムは、上記本発明の情報分類装置と、通信回線を介して流通する前記分類対象テキストを前記情報分類装置に取り込むテキスト入力手段とを備えたことを特徴とする。前記テキスト入力手段は、前記分類対象テキストをエージェント機能を通じて前記情報分類装置に入力するように構成することが望ましい。

【００１５】

【発明の実施の形態】以下、図面を参照して本発明における実施の形態を詳細に説明する。

（第１実施形態）図１は、本実施形態による情報分類装置の一実施形態を示す機能ブロック図である。本実施形態の情報分類装置１は、スタンドアロン型コンピュータ装置の内部あるいは外部記憶装置に構築される文書データベース１７と、上記コンピュータ装置が所定のプログラムを読み込んで実行することにより形成される、テキスト入力部１１、単語処理部１２、ベクトル処理部１３、学習特徴ベクトル集合ファイル１４、類似度処理部１５、カテゴリ決定部１６、を備えて構成される。

【００１６】なお、上記プログラムは、通常、コンピュータ装置の内部記憶装置あるいは外部記憶装置に格納され、随時読み取られて実行されるようになっているが、コンピュータ装置とは分離可能な記録媒体、例えばＣＤ－ＲＯＭやＦＤ等の可搬性記録媒体、あるいは当該コンピュータ装置と構内ネットワークに接続されたプログラムサーバ等に格納され、使用時に上記内部記憶装置または外部記憶装置にインストールされて随時実行に供されるものであってもよい。

【００１７】文書データベース１７は、電子化された複数の文書データ（以下、テキスト）が蓄積されるものである。このテキスト群は、予め蓄積された学習用のテキスト群（以下、学習テキスト）と、当該学習テキストに対して新規に分類対象となる１または複数のテキスト（以下、分類対象テキスト）の分類結果とが蓄積されるように構成されている。

【００１８】また、この学習テキストには、予めＮ種類のカテゴリＣ１、Ｃ２、…、ＣＮのいずれかがテキスト毎に１または複数付与されているものとしている。カテゴリが付与された学習テキストは単語処理部１２に入力される。

【００１９】テキスト入力部１１は、図示しない入力手段により、分類対象テキストの入力を受け付けて単語処理部１２への入力を行うものである。単語処理部１２は、入力されたテキストに対して所定の形態素解析を施して単語の抽出を行うとともに、抽出された複数の単語に対して、各々、重要度を付与するものである。重要度が付与された単語群は、特徴ベクトル処理部１３に入力される。なお、重要度の付与の仕方については後述する。

【００２０】ベクトル処理部１３は、単語処理部１２で付与された重要度を要素としてカテゴリ毎の特徴ベクトルまたは特徴ベクトル集合を抽出するものである。学習テキストから抽出された場合の特徴ベクトル集合（以

下、学習特徴ベクトル集合）は、学習特徴ベクトル集合ファイル14に入力されて保持され、分類対象テキストから抽出された特徴ベクトルは類似度処理部15に入力されるようになっている。

【0021】類似度処理部15は、分類対象テキストに対応する特徴ベクトルと、学習特徴ベクトル集合ファイル14に対応する特徴ベクトル集合とに基づいて、分類対象テキストの学習テキストに対する類似度をカテゴリ毎に算出するものである。算出された類似度は、カテゴリ決定部16に入力される。なお、類似度算出処理については後述する。

【0022】カテゴリ決定部16は、算出されたカテゴリ毎の類似度に基づいて分類対象テキストに付与すべきカテゴリを決定するものである。このカテゴリ決定部16は、例えば類似度が最大となるものから順次図示しないディスプレイ装置等を通じて利用者に提示し、この提示に基づいて利用者から特定されたカテゴリを分類対象テキストに付与すべきカテゴリとして決定するように構成される。このようにすれば、利用者等が必要とする情報に対して漠然としたイメージしか有していない場合であっても、類似度が高い方から低い方へ順に探索することで、必要な情報を容易に取得することが可能となる。カテゴリ決定部16は、また、決定されたカテゴリを分類対象テキストに付与して文書データベース17に送出するように構成される。これにより、文書データベース17は、分類対象テキストをカテゴリ毎に蓄積できるようになる。

【0023】次に、本実施形態の情報分類装置1を用いた情報分類方法を、学習テキスト及び分類対象テキストにおける重要度の付与、特徴ベクトルの作成、及び類似度の判定の処理を中心に説明する。単語処理部12では、まず、学習テキストに出現する複数の単語tk（1≦k≦M）を抽出し、カテゴリCi（1≦i≦N）に属する学習テキストにおける単語tkの出現件数df（tk，Ci）を算出する。この出現件数の算出は、抽出されたすべての単語t1，t2，…，tMに対応する出現件数df（t1，Ci），df（t2，Ci），…，df（tM，Ci）を各々算出するものである。

【0024】ここで、出現件数dfの大きい単語群は、必ずしもカテゴリにおける重要な単語のみとなるものではなく、前述のように特徴語と一般語とが混在しているという問題がある。具体的には、特徴語は特定のカテゴリでのみ高い出現件数を表すのに対して、一般語は多くのカテゴリで共通して高い出現件数を表すものと考えられる。そこで単語処理部12では、単語の一般性を判定するために、カテゴリ頻度cfを定義する。例えば、すべてのカテゴリ数Nにおいて特定の単語tkがn個のカテゴリに出現するような場合のカテゴリ頻度cf（tk）は、n（n≦N）で表される。即ち、特定の単語が出現するカテゴリ数を当該単語のカテゴリ頻度として定義することができる。このカテゴリ頻度cf（tk）が大きいほど、単語tkは、カテゴリへの依存の少ない一般的な単語として特定可能となる。

【0025】次に、単語tkのカテゴリCiにおける重要度W（tk，Ci）を、例えば、単語の出現件数df、及びカテゴリ頻度cfの逆数を利用した値icf（カテゴリ頻度係数）を用いて、以下に示す式（2）及び（3）のように定義する。

$$W(tk,Ci)=df(tk,Ci)\times icf(tk) \quad \cdots(2)$$

$$icf(tk)=\log(N/cf(tk)) \quad \cdots(3)$$

出現件数df及びカテゴリ頻度cfに基づく上記式（2）を用いることにより、出現件数dfの高い単語群における一般的な単語の重要度を低減させることができ、また、特徴語となる単語に対してより高い重要度を付与することが可能となる。図2に、単語の重要度算出を表す概念図を示す。

【0026】なお、単語の重要度は、上記式（2）以外にも、例えば、単語の出現頻度tfをさらに乗算する等、従来手法により利用されているパラメータとの融合により算出するように定義することもできる。

【0027】図3は、学習テキストに対応する特徴ベクトルの抽出手順説明図である。学習テキストにおけるカテゴリCiの特徴ベクトルpiは、具体的には、上記式（2）で定義した単語の重要度を各要素として、以下に示す式（4）で算出することができる。

$$pi=(W(t1,Ci),W(t2,Ci),\cdots,W(tM,Ci)) \quad \cdots(4)$$

【0028】ベクトル処理部13では、上記式（4）に基づいて、すべてのカテゴリC1、C2、…、CNについての特徴ベクトルp1，p2，…，pNを、出現件数df及びカテゴリ頻度cfに基づいて各々算出する（ステップS101～102）。これらのカテゴリ別の特徴ベクトルから成る集合、即ち学習特徴ベクトル集合は、学習特徴ベクトル集合ファイル17に保持される（ステップS103）。

【0029】一方、未分類、即ちカテゴリが付与されていない分類対象テキストTにおける特徴ベクトルqは、q＝（W'（t1），W'（t2），…，W'（tM））で算出される。ここで、W'（tk）は、分類対象テキストTにおける単語tkの重要度であり、例えば、分類対象テキストT中における単語の出現頻度tf等に基づいて算出されるものである。

【0030】この分類対象テキストTの特徴ベクトルqを用いて、類似度処理部15では、学習テキストのカテゴリに対する分類対象テキストTの類似度を算出する。この類似度は、例えば、従来手法で採用されている公知のベクトル間の内積を利用した以下の式（5）により算出することができる。

【0031】

【数1】

9

$$d(p_i, q) = \frac{p_i \cdot q}{|p_i| \cdot |q|} \quad \cdots (5)$$

【００３２】上記式（５）における「ｄ（ｐｉ，ｑ）」は、両特徴ベクトルのなす角の余弦を表しており、その値は、「－１≦ｄ（ｐｉ，ｑ）≦１」の範囲となる。この余弦ｄ（ｐｉ，ｑ）が大きいほど両特徴ベクトルの指す方向が近い、換言すれば、分類対象テキストＴがカテゴリＣｉに属する可能性が高いことを意味する。この余弦ｄ（ｐｉ，ｑ）が即ち類似度となるものであり、カテゴリ決定部１６では、分類対象テキストＴと類似度が高いと判定されるカテゴリから所定の順で分類先のカテゴリを決定する。

【００３３】図４は、分類対象テキストの分類処理の手順説明図である。なお、ここでは、学習テキストにおける学習特徴ベクトル集合は既に抽出済みであり、学習特徴ベクトル集合ファイル１４に保持されているものとする。

【００３４】分類対象テキストはテキスト入力部１１を介して単語処理部１２に入力され、単語が抽出される。そして、抽出された各単語の当該テキストにおける出現頻度と、出現頻度に基づいた重要度とが算出される。ベクトル処理部１３では、算出された各単語の重要度を要素として、分類対象テキストの特徴ベクトルｑを抽出する（ステップＳ２０１）。なお、分類対象テキストが複数の場合には、テキスト毎に特徴ベクトルｑが抽出される。類似度処理部１５は、分類対象テキストの特徴ベクトルｑと学習特徴ベクトル集合ファイル１４中の各特徴ベクトルｐｉとの類似度Ｄｉ（＝ｄ（ベクトルｐｉ，ベクトルｑ））を、すべてのカテゴリについて各々算出する（ステップＳ２０２～２０３）。

【００３５】類似度Ｄｉが算出された後、カテゴリ決定部１６は、各類似度を算出値の大きさで降順に整列し（ステップＳ２０４）、当該算出値が最大となるものから所定数を選択して当該算出値に係るカテゴリ群を分類対象テキストの属するカテゴリ候補として決定する。当該算出値が所定範囲内となるカテゴリ群を当該分類対象テキストに付与すべきカテゴリ候補とするようにしても良い。これにより分類対象テキストは、当該カテゴリで分類され（ステップＳ２０５）、文書データーベース１７に蓄積される。なお、ステップＳ２０４～２０５におけるカテゴリの決定は、類似度の算出値の大きさに着目したものであるが、この例に限定することなく、カテゴリ決定に係る閾値を適宜設定して、決定すべきカテゴリを絞り込むように構成することも可能である。

【００３６】このように、本実施形態の情報分類装置１では、学習テキストにおける単語の重要度を決定する際に、出現件数及びカテゴリ頻度（またはカテゴリ頻度係数）を用いるようにしたので、カテゴリの特徴語となる単語の候補を容易に選択できるようになった。

【００３７】また、すべてのカテゴリに出現する単語の

10

割合を重要度に反映させるようにしたので、出現件数の高い単語群における一般語の重要度を低減させ、一般語よりも高い重要度を特徴語に対して付与することができるようになった。これにより、学習特徴ベクトルの品質及び分類精度が大幅に向上した。

【００３８】（第２実施形態）本発明は、インタネット等の公衆網を介して流通する大量の電子化情報に対して自動的な分類処理を行うシステム、例えば、上記情報分類装置として機能するところの情報分類サーバ、情報取得装置として機能するところのクライアント、を配備した情報分類システムの形態での実施も可能である。この場合の情報分類サーバは、例えば、インタネット環境上における複数の大規模なデータベースに対するサーチエンジンとして位置付けられる。

【００３９】この情報分類サーバは、第１実施形態の情報分類装置１と同様、コンピュータ装置の内部あるいは外部記憶装置に、上記文書データベース１７と同一のデータベースを構築し、公衆網を介してクライアントと通信を行う通信制御部、を具備するとともに、上記情報分類装置１と同様の機能ブロック、テキスト入力部１１、単語処理部１２、特徴ベクトル処理部１３、学習特徴ベクトル集合ファイル１４、類似度処理部１５、カテゴリ決定部１６、を具備して構成される（符号は図１に従っている）。

【００４０】この情報分類サーバが上記情報分類装置１と相違する点は、通信制御を行う公知の通信制御部を具備する点であり、この通信制御部を介して流通する電子化情報群をテキスト入力部１１に入力するとともに、クライアントからの分類要求を受けるように構成する。この分類要求には、例えば、分類対象となる電子化情報を識別するための情報等を用いれば良い。分類結果も同様に、通信制御部を介してクライアントに対して送信を行うように構成することで代替が可能であり、上記情報分類装置１と同等の効果を得ることができる。この場合の分類結果としては、例えば、対象となるテキストの属するカテゴリを用いれば良い。

【００４１】また、情報分類サーバへのテキスト手段として、インタネット環境下におけるエージェント機能を用いることにより、流通する大量の電子化情報群に対して自動的な情報分類及び管理を行うことができるシステム構築が可能となる。従って、例えばクライアント側の利用者等が必要とするテキストに対して漠然としたイメージしか有していない場合であっても、テキストの分類に係る上位レベルから下位レベルへ順次分類処理を施し、その経過を辿っていくことにより、必要な情報を容易に取得することが可能となる。

【００４２】

【発明の効果】以上の説明から明らかなように、本発明によれば、学習特徴ベクトルを明確に表現できるので、高精度の分類が可能となる。また、学習テキストにおけ

11　　12

る既存の分類体系に則した本発明の分類処理を自動的に行うことにより、利用者等が必要とする情報を容易に検索して活用することが可能となる。さらに、本発明を情報検索システム等に適合させた場合には、検索処理の効率及び実用性が格段に向上するシステムの提供が可能となる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る情報分類装置における機能ブロック図。

【図２】単語の重要度算出を表す概念図。

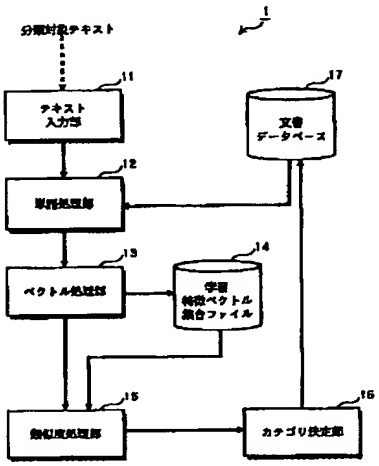
【図３】学習特徴ベクトル集合作成における処理手順図。

【図４】分類処理における処理手順図。

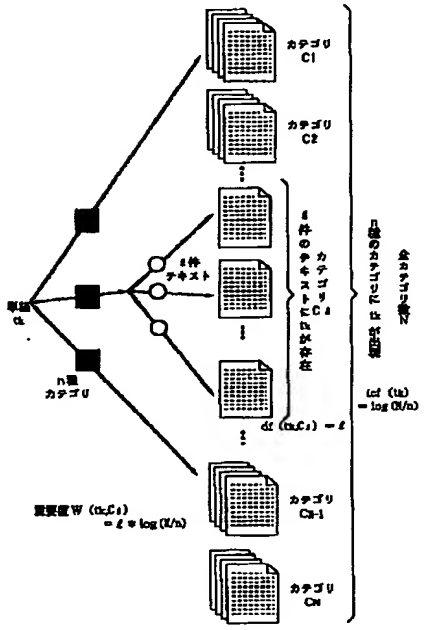
【符号の説明】

１　情報分類装置
１１　テキスト入力部
１２　単語処理部
１３　特徴ベクトル処理部
１４　学習特徴ベクトル集合ファイル
１５　類似度処理部
１６　カテゴリ決定部
１７　文書データベース

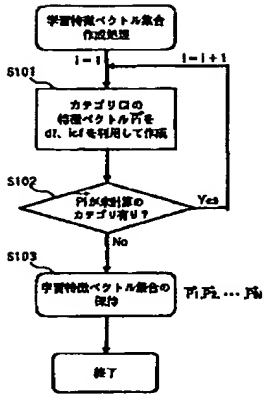
【図１】



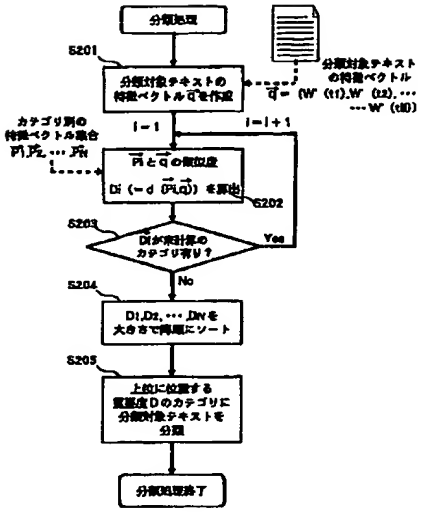
【図２】



【図３】


学習特徴ベクトル集合
作成処理
I＝1
I＝I＋1
S101
カテゴリIの
特徴ベクトルPIを
df、idfを利用して作成
S102
PIが未計算の
カテゴリ有り？
Yes
No
S103
学習特徴ベクトル集合の
保持
P1,P2,…PN
終了

【図４】


分類処理
S201
分類対象テキストの
特徴ベクトルQを作成
分類対象テキスト
の特徴ベクトル
Q＝（W（t1),W（t2),…
…W（tn））
カテゴリ別の
特徴ベクトル集合
P1,P2,…PN
I＝1
I＝I＋1
PIとQの類似度
DI（＝d（PI,Q））を算出
S202
S203
DIが未計算の
カテゴリ有り？
Yes
No
S204
D1,D2,…DNを
大きさで降順にソート
S205
上位に位置する
重要度Dのカテゴリに
分類対象テキストを
分類
分類処理終了